SUPPLEMENTARY RESULTS

The ENCODE4 microRNA-seq dataset. We sequenced 254 human samples with microRNA-seq, which specifically captures mature (21-25 bp) microRNAs. We mapped our microRNA-seq data to pre-microRNA sequences in GENCODE v29. We detected 1,130 microRNAs at CPM >= 2 across the full dataset (Fig. S2a). Overall, microRNAs are more sample-specific than known genes in both short and long-read RNA-seq (Fig. S2a, Fig. 1c, f, Fig. S3b). PCA of all microRNA samples shows separation of brain samples from other tissues and cell lines by PC1, while cell lines are separated by both PC1 and PC2 (Fig. S2b). Comparison of microRNA detection between tissue types and cell lines reveals that brain samples have the least diversity when compared to the full set of GENCODE v29 microRNAs (Fig. S2c), yet the most microRNAs expressed at >=2 CPM (Fig. S2d). This indicates that a core set of microRNAs are expressed in the brain. Their high tissue-specific expression may be driving the clustering of brain samples apart from non-brain tissues and cell lines, which overlap slightly (Fig. S2b). Comparison of the overlap of detected microRNAs across the sample biotypes reveals that more microRNAs overlap between brain and cell lines than between brain and non-brain tissues (Fig. S2e). Of the 80 shared microRNAs between brain and cell lines, most (53) are expressed in neuronal and glial derived cells.

Machine learning models predict the support for long-read TSS peaks by other TSS-annotating assays and in a cross-cell type manner. We sought to identify a set of high-confidence TSS regions from our observed LR-RNA-seq TSSs using multiple orthogonal TSS assays such as RAMPAGE and CAGE[40,41]. However, matching data from external assays is only available for a few samples, such as our ENCODE tier 1 cell lines GM12878 and K562. Therefore, we wanted to predict the external support for our observed LR-RNA-seq TSSs. The majority of our observed TSS regions are supported by these external assays (Fig. S6a-b, Fig. 2b, Fig. S5a). We used a simple logistic regression model that incorporates expression, DNase-Hypersensitivity (DHS)[42], and length of our LR-RNA-seq observed TSSs (Supplementary methods, Fig. S6c-d). Models trained and tested on one experiment each from GM12878 and K562 TSS regions were able to predict whether an LR-RNA-seq TSS region was also supported by RAMPAGE or CAGE assays, with AUROC values as high as 0.95 for Cerberus and 0.98 for LAPA-annotated peaks in the same cell type (Fig. S6e), which is expected given that LAPA regions are narrower than Cerberus regions. Models trained on one cell type can also be used to predict the RAMPAGE or CAGE support in another cell type, in a cross-cell type manner (Fig. S6f). This approach may be used to define a set of high-confidence TSS regions from LR-RNA-seq that would also be supported by RAMPAGE or CAGE where neither RAMPAGE nor CAGE data are available in the cell type of interest. This demonstrates that TSSs derived from LR-RNA-seq serve as a reasonable stand-in for CAGE and RAMPAGE, with the added benefit that LR-RNA-seq profiles both ends and the exon structure of transcripts at the same time.

Applying Cerberus to the human ENCODE4 LR-RNA-seq dataset leads to the largest number of detected alternative splicing events to date. We compared the detection of alternative splicing (AS) events in our dataset with a recent LR-RNA-seq transcriptome published by the GTEx consortium[23]. We ran SUPPA2[43] on the observed LR-RNA-seq transcripts and obtained, for every gene and type of local AS event, a list of AS transcripts. We found a considerably larger number of AS transcripts compared to those reported in the GTEx LR-RNA-seq catalog. We observed a higher proportion of novel AS transcripts defined by EC compared to TSS and TES (Fig. S8a), albeit lower than those reported by GTEx. This is likely due to the fact that our novel transcripts are defined with respect to a more recent GENCODE version (v40) than the one used by the GTEx study (v26). In support of this, we found that the majority of our observed ENCODE LR-RNA-seq transcripts, both known and novel, are missing in the GTEx catalog (Fig. S8b). On the other hand, although most of the GTEx novel transcripts are not reported in the ENCODE4 catalog, they represent combinations of already annotated splice junctions (NIC). From a methodological perspective, we also found that Cerberus accounts for a larger variety of AS events related to TSSs and TESs ($0.25 < PSI < 0.75$) compared to SUPPA2 (Fig. S8c). Altogether, this indicates that the ENCODE4 LR-RNA-seq catalog provides the largest set of novel and annotated AS events in the human transcriptome available to date.

SUPPLEMENTARY METHODS

B6/Cast mouse tissue collection. Mouse tissues were harvested from C57BL/6J (RRID:IMSR_JAX:000664) x CAST/EiJ (RRID:IMSR_JAX:000928) F1 animals across 7 postnatal day (PND) or postnatal month (PNM) timepoints: PND4, PND10, PND14, PND25, PND36, 2 months and 18-20 months. Tissues were flash frozen in liquid nitrogen and stored at -80C prior to processing.

Preprocessing short-read RNA-seq data and data availability. All short-read RNA-seq data was preprocessed according to the details on the ENCODE portal. Gene quantification of 548 short RNA-seq datasets were downloaded from the ENCODE portal using this cart (https://www.encodeproject.org/carts/4ea7a43f-e564-4656-a0de-b09c92215e52/), then TPM values for polyA genes were extracted from each of them.

Preprocessing microRNA-seq data and data availability. Quantification of 254 microRNA-seq datasets using GENCODE GRCh38 V29 annotations were downloaded from the ENCODE portal using this cart (https://www.encodeproject.org/carts/human_mirna/). Counts were concatenated across all datasets and converted to CPM for downstream analyses.

Preprocessing LR-RNA-seq data and data availability. All LR-RNA-seq data was preprocessed according to the details on the ENCODE portal. Input and output files, including the final Cerberus GTFs, gene triplets, and transcript triplets, are available at the following accessions:

- Human: ENCSR957LMA
- Mouse: ENCSR110KDI

Raw data are available at the following links:

- Human: https://www.encodeproject.org/carts/829d339c-913c-4773-8001-80130796a367/
- Mouse: https://www.encodeproject.org/carts/55367842-f225-45cf-bfbe-5ba5e4182768/

Human / Mouse LR-RNA-seq annotation with TALON and LAPA. Mapped LR-RNA-seq BAMs were obtained from the ENCODE portal using the above cart links for human and mouse respectively. Reads were annotated with their 3' end A content using the **`talon_label_reads`** module and hg38 / mm10. Reads were annotated using talon with reference annotation GENCODE v29 / vM21. Output transcripts were filtered for reproducibility of 5 reads across 2 libraries, and for reads that had fewer than 50% A nucleotides in the last 20 bp of the 3' end to remove artifacts of internal priming using the **`talon_filter_transcripts`** command. Unfiltered and filtered transcript abundance matrices were obtained using the **`talon_abundance`** command. A filtered GTF was obtained using the **`talon_create_GTF`** command. From the unfiltered TALON abundance, counts of each gene were computed by summing up counts for each transcript per gene.
We ran LAPA on the bam files mentioned above to create TSS and TES clusters from LR-RNA-seq. If the bam files had replicates, we filtered clusters by choosing a cutoff that ensures a 95% replication rate. Samples without replicates were filtered with a median cutoff of replicated clusters. Using those TSS and TES clusters and the read_annot created by TALON, we corrected TSSs and TESs of the filtered TALON GTF file. During the correction, new transcript isoforms were created if the same exon junction chain mapped to multiple start and end sites.

Gene rank analysis. For detected (>= 1 TPM in any library) polyA genes in the human LR-RNA-seq dataset, we ranked the genes in each library according to their expression (1 = most highly expressed) and plotted the genes at specific ranks for each library by their TPM, split by cell line and tissue derived libraries. For statistical testing between the cell line and tissue groups, we performed a Wilcoxon rank-sum test with p-value thresholds P > 0.05; *P <= 0.05, **P <= 0.01, ***P <= 0.001, ****P <= 0.0001.

Novel gene analysis. For novel genes in both human and mouse, we first filtered our novel TALON transcripts for those that passed the filters previously described (5 reads in at least 2 libraries and <50% A nucleotides in the last 20bp of the 3' end). We then selected only the transcripts that passed this filter that belonged to novel intergenic genes and that had at least one spliced (i.e. more than one exon) transcript isoform expressed >= 1 TPM. To make an analogous comparison to our annotated genes, we performed the same filtering on our TALON transcripts with the exception of requiring the transcripts to be from annotated polyA genes rather than from novel intergenic genes.

Cerberus overview.

Obtaining annotated TSS / TES regions from GTFs. Given a GTF, **`cerberus gtf_to_bed`** (Fig. S4a) will extract the single base pair TSS and TES coordinates and extend them by *n* bp on either side. Regions within *m* bp of one another are merged. Each unique combination of coordinates, strand, and gene is recorded in BED format.

Obtaining annotated exon junction chains from GTFs. Given a GTF, **`cerberus gtf_to_ics`** will extract each unique combination of intron coordinates, strand, and gene and record them in a tab-separated format (Fig. S4a).

Assigning triplet features numbers. As part of both **`cerberus gtf_to_ends`** and **`gtf_to_ics`** (Fig. S4a), Cerberus numbers triplet features based on their annotation status within the reference GTF, if any. For triplet features derived from these GTFs, each TSS, EC, and TES is numbered from 1 to *n* within each gene based on the annotation status of the transcript they were derived from. Transcripts are first ordered by MANE status, then APPRIS[44] principal status, and finally whether the transcript comes from the GENCODE basic set. The result is that triplet features from MANE transcripts are always numbered 1, and lower triplet feature numbers within a gene correspond to transcripts with more importance as determined by GENCODE.

Merging TSS and TES regions across multiple BED files. For each BED file input **`cerberus agg_ends`** (Fig. S4a) takes a boolean argument for whether the regions should be used to initialize new TSS / TES regions, a boolean argument for whether the regions should be considered reference regions, and a name for each BED file source. BED files without a gene identifier cannot be used to initialize regions. For the first BED file, Cerberus creates a set of reference regions and uses the triplet feature numbers that were previously assigned by Cerberus to name each TSS or TES. The first BED file must have gene IDs and must be used to initialize the regions. For each subsequent BED file, in order, Cerberus determines which new regions are within m bp of a region already in the reference. These regions are added as sources of support for the already-existing regions, but do not extend the boundaries of existing regions in order to combat growing regions as more data is added. If a region is not within m bp of an existing region and the initialize regions option is turned on, the new region is added as a new region in the reference set. After all new regions have been added, triplet feature numbers are computed by ordering the features within each gene based on the number assigned by Cerberus in a previous step and then incrementing the preexisting Cerberus reference maximum number. BED files that are not used to initialize new regions will only ever be added as additional forms of support for each region already in the reference.

Merging ECs across multiple EC files. For each EC file, **`cerberus agg_ics`** (Fig. S4a) takes a boolean argument for whether the ECs should be used as a reference and a source name. For the first EC file, Cerberus creates a reference set of ECs and uses the EC numbers that were determined using **`cerberus gtf_to_ics`**. For each subsequent EC file, Cerberus finds ECs that are not already in the Cerberus reference set, orders the new ECs by their numbers from **`cerberus gtf_to_ics`**, creates new numbers for each EC by, in order, assigning them numbers by incrementing from the maximum existing number for a gene from the reference.

Creating a Cerberus reference. After generating an aggregated TSS, EC, and TES file, **`cerberus write_reference`** (Fig. S4a) will write all three tables in a Cerberus reference h5 format, a well-supported and commonly used data structure that can store multiple tables.

Updating a GTFs and counts matrices with a Cerberus annotation. After each transcript from a transcriptome has been assigned a transcript triplet, the corresponding GTF and counts matrix from the transcriptome can be updated to use the new transcript identifier using **`cerberus replace_gtf_ids`** and **`cerberus replace_ab_ids`** (Fig. S4b). Cerberus will replace the transcript ids with the transcript triplets and, if requested, merge transcripts that are assigned duplicate transcript triplets, summing the counts in the case of the counts matrix.

Gene triplet and gene structure simplex coordinate computations. Following transcriptome annotation, gene triplets can be calculated for different sets of annotated transcripts using Cerberus' Python API and the **`CerberusAnnotation data structure`**. Regardless of the input set, Cerberus computes the gene triplets by counting the number of unique TSSs, ECs, and TESs used across a set of transcripts (Fig. 3a, Fig. S9). This calculation can be done without any filtering using the **`CerberusAnnotation.get_source_triplets()`** function, which computes the number of TSSs, ECs, and TESs used across each transcriptome annotated by Cerberus. **`CerberusAnnotation.get_expressed_triplets()`** will calculate the gene triplets for individual samples based on the subset of transcripts that are expressed in each sample and can optionally use a table of transcript / sample combinations to determine which transcripts are used in each sample. Finally, **`CerberusAnnotation.get_subset_triplets()`** simply takes in a list of transcripts to compute a gene triplet for the entire input set. In all cases, the number of transcripts used to calculate the gene triplet is also recorded. After computing the gene triplets, the EC count is converted to the splicing ratio. To generate the gene structure simplex coordinates, the sum of the number of TSSs, splicing ratio, and number of TESs is normalized such that they sum to one (Fig. 3a-b, Fig. S9).

Additionally, the sector assignments are generated for each gene triplet. Genes with a TSS simplex coordinate >0.5 are TSS-high, those with a TES simplex coordinate >0.5 are TES-high, and those with a splicing ratio simplex coordinate >0.5 are splicing-high.Genes where all three simplex coordinates <= 0.5 are mixed, and genes with just one transcript are in the simple sector. An important note is that mixed genes can have the same coordinates as a simple gene. To this end, when calculating gene triplets, the number of transcripts used to generate the triplet is also recorded and used to separate out the simple from the mixed genes (Fig. 3a-b, Fig. S9).

Computing gene triplet centroids. Given a set of gene triplets, the centroid is computed by averaging each gene structure simplex coordinate. The resulting coordinate retains the property that it sums to one (Fig. S9).

Computing distances in the gene structure simplex. We compute the distance between any two points on the gene structure simplex as the Jensen-Shannon distance (Fig. S9). Jensen-Shannon distance is a metric on probability distributions[45]. For a given pair of gene structure simplex coordinates, the Jensen-Shannon distance is computed in Cerberus using Scipy[46] with the **`scipy.spatial.distance.jensenshannon`** function.

Cerberus processing of human ENCODE4 LR-RNA-seq dataset.

Obtaining annotated TSS / TES regions from GTFs. The GTF files from GENCODE v40, GENCODE v29, the LAPA output GTF representing the human ENCODE LR-RNA-seq dataset, and the GTEx LR-RNA-seq GTF were used to obtain TSS and TES regions associated with each transcript using **`cerberus gtf_to_ends`**. For each GTF, the single base pair TSS and TES coordinates were extracted and extended 50 bp on either side, and regions within 50 bp of one another were merged. Each unique combination of coordinates, strand, and gene were recorded.

Obtaining external TSS / TES regions. External datasets used to support TSSs were obtained from the ENCODE CAGE and RAMPAGE data, FANTOM CAGE data4, and ENCODE PLS, pELS, and dELS cCREs. External datasets used to support TESs were obtained from ENCODE PAS-seq data, and the PolyA Atlas[18]. Each file was downloaded in BED format and converted to the BED format required for Cerberus.

Obtaining annotated exon junction chains from GTFs . The GTF files from GENCODE v40, GENCODE v29, from the human ENCODE LR-RNA-seq output GTF, and the GTEx LR-RNA-seq GTF were used to obtain exon junction chains from each transcript using **`cerberus gtf_to_ics`**. Each unique combination of intron coordinates, strand, and gene were recorded.

Creating a set of reference triplet features. To create a consensus reference set of triplet features, **`cerberus agg_ends`** and **`cerberus agg_ics`** (Fig. S4a) were run on the aforementioned TSS, EC, and TES sets, with m=20 for the TSSs and TESs. The triplet features from GENCODE v40 and v29 were used as reference features. For the TSSs, new regions were incorporated from GENCODE v40, v29, the human ENCODE LR-RNA-seq data, and the GTEx data, whereas the CAGE, RAMPAGE, and cCRE data were only used as forms of support for existing regions. For TESs, new regions were incorporated from GENCODE v40, v29, the human ENCODE LR-RNA-seq data, and the GTEx data, whereas the PAS-seq and PolyA Atlas regions were used as forms of support for existing regions.

Transcriptome annotation. The GTFs of the GENCODE v40, GENCODE v29, and human ENCODE LR-RNA-seq transcriptomes were annotated with **`cerberus annotate_transciptome`**, updated GTFs were generated with **`cerberus replace_gtf_ids`** with the update ends and collapse options used (Fig. S4b). For the human ENCODE LR-RNA-seq data, **`cerberus replace_ab_ids`** was also run on the filtered abundance file output from LAPA using the collapse option to generate a matching counts matrix (Fig. S4b).

Cerberus analysis of human ENCODE4 LR-RNA-seq.

Finding observed transcripts and transcripts expressed in a sample. Observed transcripts are defined as transcripts that are expressed >= 1 TPM in any given library. Observed transcripts in a specific sample are transcripts that are expressed >= 1 TPM in any library that belongs to the same sample.

Finding observed major transcripts and major transcripts in a sample. For each sample, each transcript is assigned a percent isoform (pi, 0-100) value that indicates what percentage of the gene's expression is derived from said transcript using Swan[47]. Transcripts for a gene are then ranked by pi value. In order from the highest pi value transcript to the lowest pi value transcript, transcripts are added to the major transcript set until the cumulative pi value of the set is >90, yielding the sample-level major transcript set. The observed major transcripts for the entire dataset is computed by taking the union of all major transcripts across all samples. In both cases, transcripts are limited to those that have passed the observed and sample-level observed transcripts as defined above.

Gene triplet computations. Gene triplets were calculated for the following sets of transcripts, all just using polyA genes:

- All transcripts from annotated GENCODE v40 genes (v40)
- All observed transcripts (observed)
- All observed major transcripts (observed major)
- Detected transcripts in each sample (sample-level)
- Detected major transcripts in each sample (sample-level major)
- All observed transcripts in the dataset from samples that match the mouse samples (mouse match)
- All observed major transcripts in the dataset from samples that match the mouse samples (mouse match major)

Transcriptional diversity by gene biotype comparison. Using the gene triplets table, we found the gene / sample combination where each polyA gene is most highly expressed and recorded the gene TPM and number of transcripts from that gene in that sample. We then split each gene into its biotype category (protein coding, lncRNA, or pseudogene) and into a TPM bin (lowly expressed, 1-10 TPM; medium expressed, 10-100 TPM; and highly expressed, 100-max TPM).

Gene structure simplex distances computed. We computed the follow pairwise distances between simplex points:

- Sample-level gene triplet vs. the centroid for the sample-level gene triplets
- Observed gene triplet vs. the centroid for the sample-level gene triplets for each gene with at least 2 transcripts
- Observed gene triplet vs. observed major gene triplet

Each set of distances was computed using only protein coding genes. Z-scores were also computed for each comparison.

Comparing sample-level to observed gene triplets. The number of transcripts, TSSs, ECs, and TESs was calculated for each gene globally (i.e. transcripts or triplet features / gene) and for each sample (i.e. transcripts or triplet features / gene / sample). For transcripts, TSSs, ECs, and TESs separately, a two-sided KS test was performed using Scipy's **`stats.kstest`** function to assess statistical differences between the global and sample-level transcripts or triplet features per gene distributions.

Calling predominant transcripts. On both the sample and library level, we called the most highly expressed transcript from a gene the predominant transcript for that gene. On the sample level, we used the mean expression of the transcript.

Predominant transcript MANE comparison. We first restricted this analysis to only consider genes which have annotated MANE transcripts. For these genes, we determined how often the predominant transcript for a given gene is the MANE transcript for a gene in each library.

Cerberus processing of mouse ENCODE4 LR-RNA-seq dataset.

Obtaining annotated TSS / TES regions from GTFs. The GTF files from GENCODE vM25, GENCODE vM21, and from the LAPA output GTF representing the mouse ENCODE LR-RNA-seq dataset were used to obtain TSS and TES regions associated with each transcript using **`cerberus gtf_to_ends`**. For each GTF, the single base pair TSS and TES coordinates were extracted and extended 50 bp on either side, and regions within 50 bp of one another were merged. Each unique combination of coordinates, strand, and gene were recorded.

Obtaining external TSS / TES regions. External datasets used to support TSSs were obtained from the ENCODE mouse PLS, pELS, and dELS cCREs. External datasets used to support TESs were obtained from the mouse PolyA Atlas. Each file was downloaded in BED format and converted to the BED format required for Cerberus.

Obtaining annotated exon junction chains from GTFs. The GTF files from GENCODE vM25, GENCODE vM21, and from the mouse LR-RNA-seq GTF were used to obtain exon junction chains from each transcript using **`cerberus gtf_to_ics`**. Each unique combination of intron coordinates, strand, and gene were recorded.

Creating a set of reference triplet features. To create a consensus reference set of triplet features, **`cerberus agg_ends`** and **`cerberus agg_ics`** (Fig. S4a) were run on the aforementioned TSS, EC, and TES sets, with m=20 for the TSSs and TESs. The triplet features from GENCODE vM25 and vM21 were used as reference features. New TSSs were incorporated from GENCODE vM25, vM21, the mouse ENCODE LR-RNA-seq data, whereas the cCRE data were only used as forms of support for existing regions. New TESs were incorporated from GENCODE vM25, vM21, and the mouse ENCODE LR-RNA-seq data, whereas the PolyA Atlas regions were used as forms of support for existing regions.

Transcriptome annotation. The GTFs of the GENCODE vM25, GENCODE vM21, and mouse ENCODE LR-RNA-seq transcriptomes were annotated with **`cerberus annotate_transciptome`**, updated GTFs were generated with **`cerberus replace_gtf_ids`** with the update ends and collapse options used (Fig. S4b). For the mouse ENCODE LR-RNA-seq data, **`cerberus replace_ab_ids`** was also run on the filtered abundance file output from LAPA using the collapse option to generate a matching counts matrix (Fig. S4b).

Cerberus analysis of mouse ENCODE4 LR-RNA-seq.

Finding observed transcripts and transcripts expressed in a sample. Observed transcripts are defined as transcripts that are expressed >= 1 TPM in any given library. Observed transcripts in a specific sample are transcripts that are expressed >= 1 TPM in any library that belongs to the same sample.

Finding observed major transcripts and major transcripts in a sample. For each sample, each transcript is assigned a percent isoform (pi, 0-100) value that indicates what percentage of the gene's expression is derived from said transcript using Swan[47]. Transcripts for a gene are then ranked by pi value. In order from the highest pi value transcript to the lowest pi value transcript, transcripts are added to the major transcript set until the cumulative pi value of the set is >90, yielding the sample-level major transcript set. The observed major transcripts for the entire dataset is computed by taking the union of all major transcripts across all samples. In both cases, transcripts are limited to those that have passed the observed and sample-level observed transcripts as defined above.

Gene triplet computations. Gene triplets were calculated for the following sets of transcripts; all just using polyA genes:

- All annotated GENCODE vM25 genes (vM25)
- All observed transcripts in the dataset (observed)
- All observed major transcripts in the dataset (observed major)
- Detected transcripts in each sample (sample-level)
- Detected major transcripts in each sample (sample-level major)

Calling predominant transcripts. On both the sample and library level, we called the most highly expressed transcript from a gene the predominant transcript for that gene. On the sample level, we used the mean expression of the transcript.

Human-mouse comparison. We found orthologous genes between human and mouse using this Biomart query, and subset our considered genes to those that were protein coding, expressed in both species, and were just 1:1 orthologs. We determined the sector of each gene in each species using the observed major gene triplets in mouse, and the mouse match major gene triplets in human. We counted the number of genes that have the same sector between human and mouse. Furthermore, we compared the sector of each orthologous pair of genes between species just in the matching embryonic stem cell samples (H1 in human, F121-9 in mouse) between human and mouse to verify that the trend seen overall was reproducible on a more one to one comparison. Additionally, we computed the centroids from the sample-level gene triplets from matching samples in human and all sample-level gene triplets in mouse mouse and calculated the Jensen-Shannon distances between sample-level centroids for each orthologous gene in human and mouse.

ORF and NMD prediction. We used TAMA's[48] ORF / NMD prediction pipeline with minimal changes to support our file formats. To pick one representative ORF from each transcript, we chose the ORF with the highest percent identity from BLASTP[49] to an annotated GENCODE v40 protein sequence; breaking ties by considering ORF completeness. For transcripts with no BLASTP hits to known transcripts, we picked complete ORFs; breaking ties by picking the longest ORF.

Comparing detection of AS events by SUPPA and Cerberus. We used SUPPA2 (v2.3[43]) to define alternative splicing (AS) events (A3: alternative 3' splicing; A5: alternative 5' splicing; AF: first exon; AL: last exon; IR: intron retention; SE: exon skipping; MX: mutually exclusive exons). Specifically, we generated a catalog of local AS events based on the Cerberus GTF file (function **`generateEvents`**) and used the novelties of each observed transcript to compute the proportion of novel transcripts (based on EC, TSS, or TES) out of the total set of transcripts involved in a particular type of event.

Next, we used SUPPA2 to compute the Proportion of Splicing Index (PSI) for each type of event using the observed transcript filtered expression matrix (polyA transcripts expressed >=1 TPM in at least one library; function **`psiPerEvent`**). PSI values were averaged between replicates of the same sample. We selected genes with at least one local AF or AL event, applying a threshold of $0.25 < PSI < 0.75$.

In order to compare the detection of AS events by SUPPA2 and Cerberus, we also computed triplet feature PSI values based on events identified by Cerberus by dividing the counts for any given TSS, EC, or TES by the total counts for the gene in a given sample. We selected genes with at least one local event at the TSS or TES ($0.25 < PSI < 0.75$). Next, we computed the intersection between genes showing AF (SUPPA2) and TSS (Cerberus) events, and between genes showing AL (SUPPA2) and TES (Cerberus) events.

Machine learning models for RAMPAGE and CAGE TSS prediction. RAMPAGE and CAGE TSS annotation data for GM12878 and K562 were obtained from ENCODE portal (ENCSR000AEI, ENCSR000AER, ENCSR000CJN, ENCSR000CKA). LAPA and Cerberus TSS regions derived from just one experiment each for GM12878 and K562 (ENCSR962BVU and ENCSR589FUJ respectively) were used for long-read data. Using bedtools intersect[50] a binary (0/1) label for each long-read peak was assigned depending on whether the region overlapped with at least one peak in either of the RAMPAGE or CAGE assays in the same cell type. Average DHS signal values over LR TSS peaks were calculated using UCSC **`bigWigAverageOverBed`** on GM12878 and K562 DHS-seq experiments (ENCSR000EMT, ENCSR000EOT). Test sets include long-read regions from chromosomes 2 and 3, whereas training sets include all other human chromosomes. 7 logistic regression models were trained on each long-read experiment using all the 2^3-1 combinations of peak's TPM expression,

DHS signal, and length (i.e. in R: glm(label TPM + DHS + length, type ="binomial")) where the input parameters have been log2-transformed) and the AIC values were calculated and ranked in each experiment and for each model type. The model using all 3 parameters (TPM, DHS, and length) had the lowest AIC, meaning that given the number of parameters and the observed RSS error, logit[label TPM + DHS + length] had the highest predictive power and was therefore selected. For the same cell-type prediction, a model is trained on [chr1, chr4-22, chrX] and tested on [chr2, chr3] for long-read data from the same cell type (ex: K562). In cross-cell type prediction, a model is trained and tested on two different cell lines (ex: trained on [chr1, chr4-22, chrX] of a GM12878 long-read experiment and tested on [chr2, chr3] of a K562 long-read experiment). Cerberus replicates belonging to the same experiment were combined by taking the average mean-normalized TPM values of the identical peaks across different replicates.

Data and code availability.

- Human LR-RNA-seq data / processing pipeline: https://www.encodeproject.org/annotations/ENCSR957LMA/
- Mouse LR-RNA-seq data / processing pipeline: https://www.encodeproject.org/annotations/ENCSR110KDI/
- Processing / figure generation code: https://github.com/fairliereese/paper_rnawg
- Cerberus: https://github.com/fairliereese/cerberus

SUPPLEMENTARY TABLES

- **Table S1: Human LR-RNA-seq library metadata.**
- **Table S2: Mouse LR-RNA-seq library metadata.**

SUPPLEMENTARY FIGURES

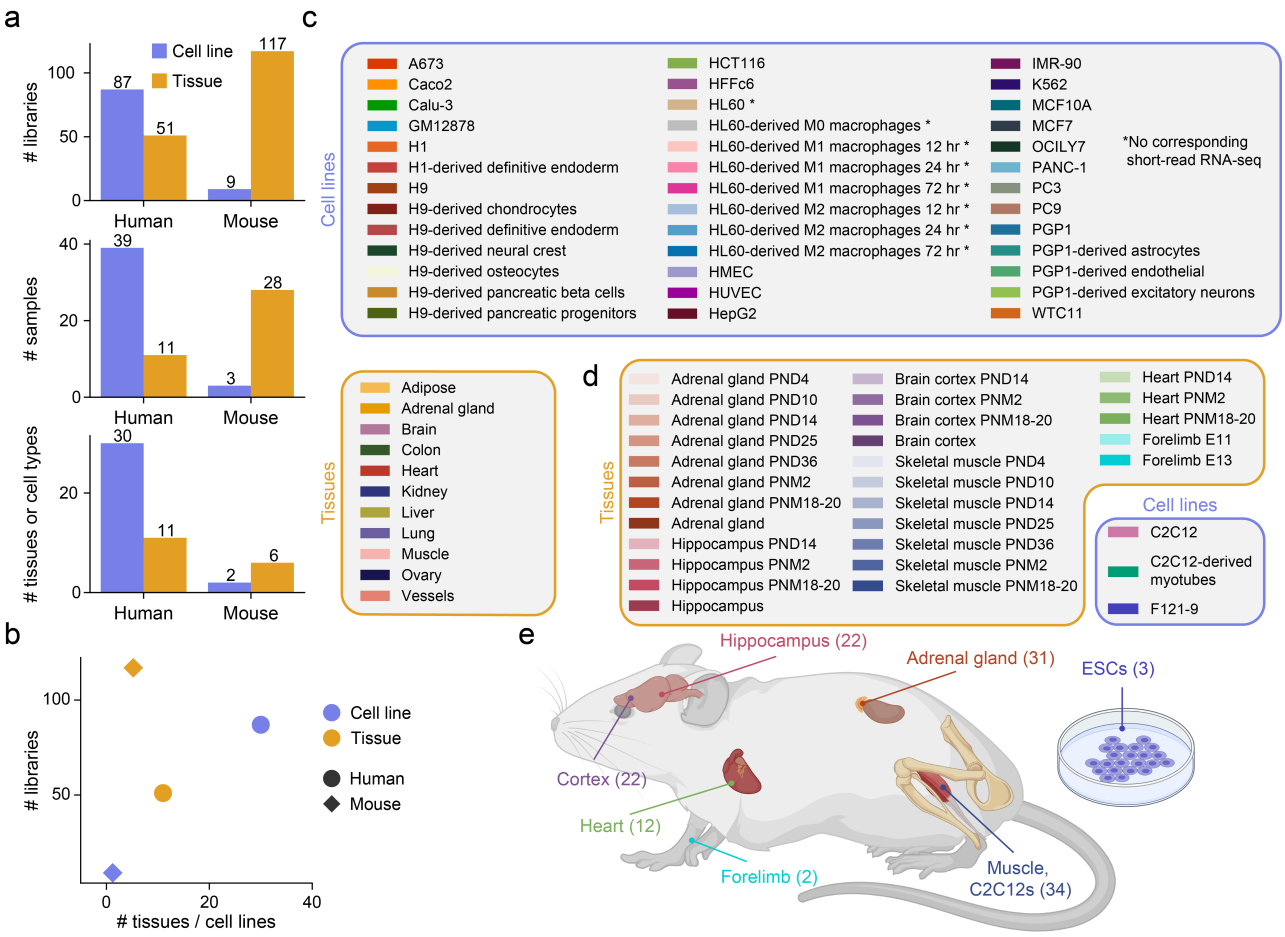

Figure S1. Overview of the ENCODE4 LR-RNA-seq dataset. a, From top to bottom, number of LR-RNA-seq libraries, samples (split by cell line / tissue identity as well as timepoint, when relevant), and unique tissues or cell types in the ENCODE LR-RNA-seq dataset split by species and tissue or cell line. **b,** Number of LR-RNA-seq libraries versus the number of tissues or cell lines assayed, split by species and cell line / tissue. **c-d,** Color legend and labels for each **c,** human sample, with samples that lack corresponding short-read RNA-seq data denoted by a star **d,** mouse sample in the LR-RNA-seq dataset; split by tissues and cell lines. **e,** Overview of the sampled tissues and number of libraries from each tissue in the ENCODE mouse LR-RNA-seq dataset.

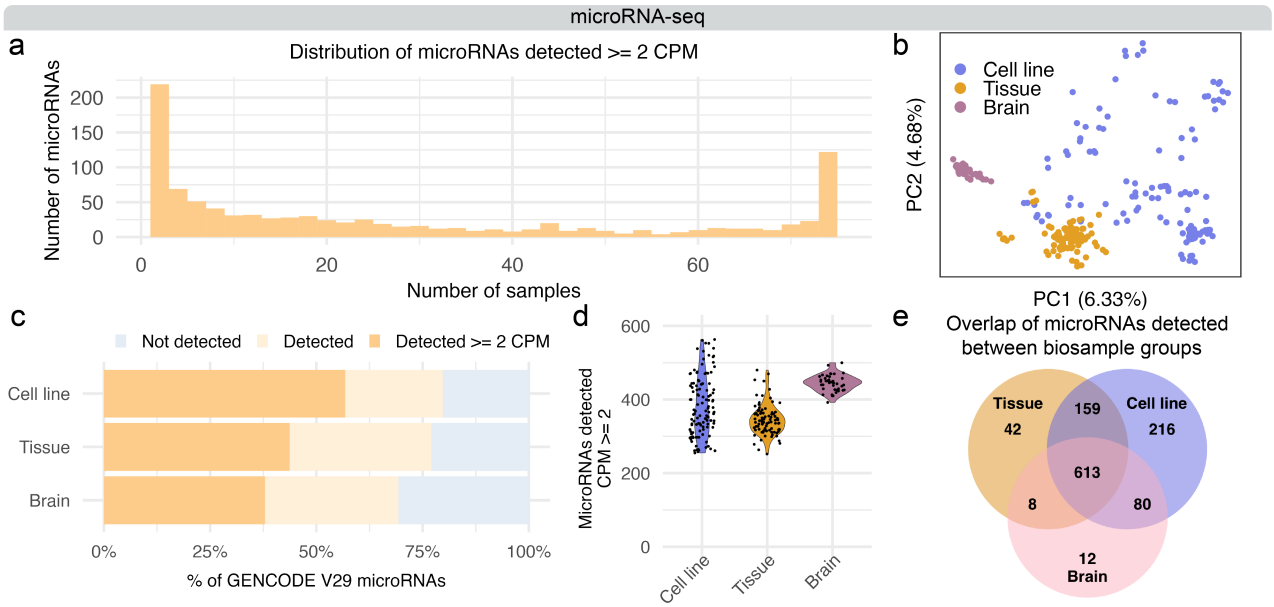

Figure S2. Overview and detection of microRNAs in the ENCODE microRNA-seq dataset. a, Distribution of GENCODE v29 mature microRNAs detected at CPM > 2 between cell lines, tissues, and brain tissue samples. **b,** PCA computed on microRNAs detected >2 CPM in each human microRNA-seq library, colored by cell line and tissue designation and by brain tissue. **c,** Percentage of GENCODE v29 microRNAs detected in at least one ENCODE human microRNA-seq library from either cell line, tissue, or brain tissue samples at > 0 CPM and > 2 CPM. **d,** Number of samples in which each GENCODE v29 microRNA is detected at > 2 CPM in the ENCODE human microRNA-seq dataset. **e,** Overlap of detected (> 2 CPM) microRNAs in at least one library derived from cell line, tissue, or brain tissue.

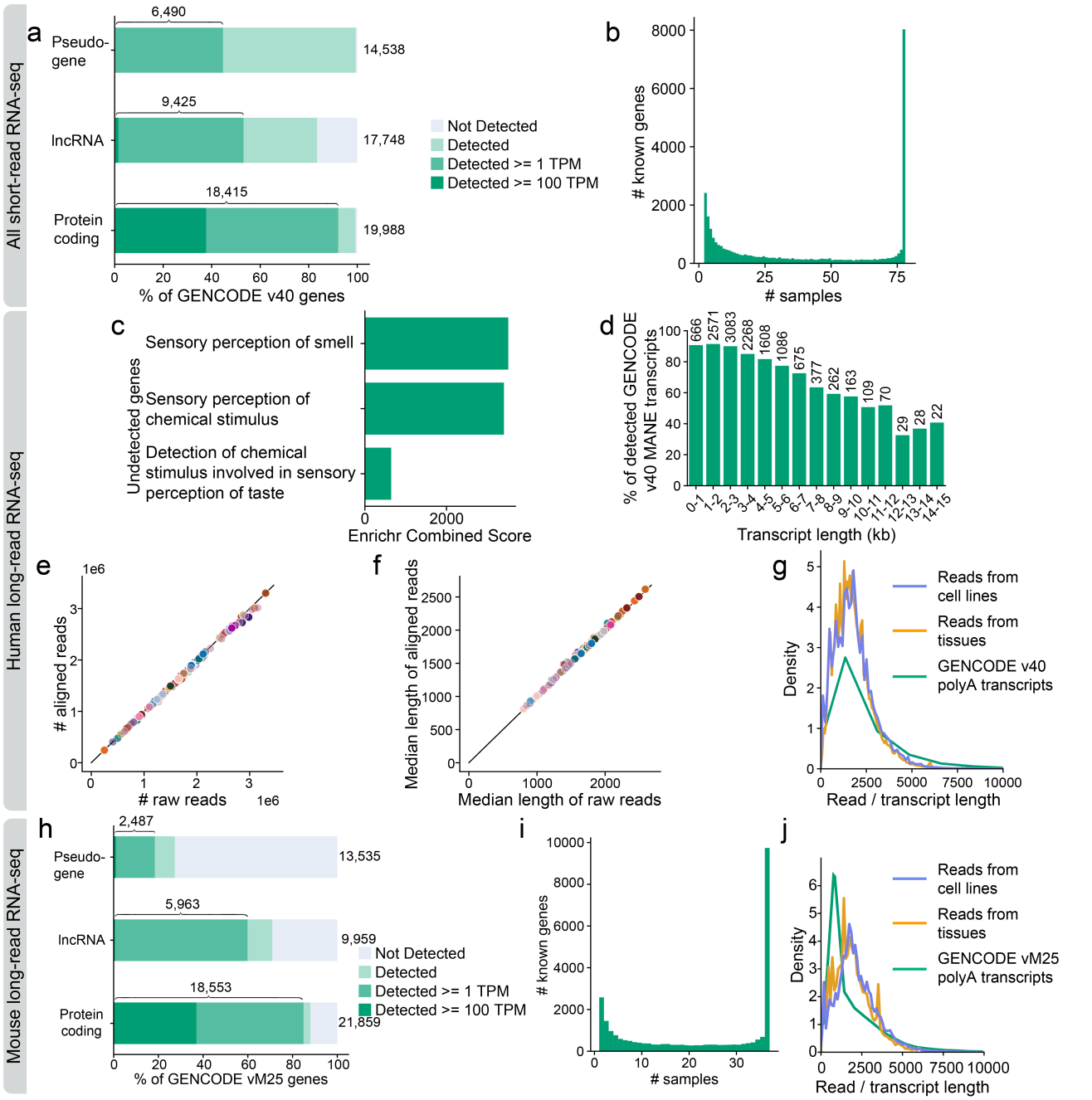

Figure S3. Gene detection from short-read RNA-seq; gene detection, read length and alignment QC in both human and mouse LR-RNA-seq. a, Percentage of GENCODE v40 polyA genes by gene biotype detected in at least one ENCODE short-read RNA-seq library from all samples at >0 TPM, >=1 TPM, and >=100 TPM. **b,** Number of samples in which each GENCODE v40 gene is detected >= 1 TPM in the ENCODE short-read RNA-seq dataset in all samples. **c,** Top 3 biological process GO terms from GENCODE v40 protein coding genes that were not detected in the human LR-RNA-seq dataset. **d,** Percentage (y-axis) and number (top of each bar) of GENCODE v40 MANE transcripts we detect binned by transcript length. Restricted to expressed genes that we detect >= 10 TPM in at least one library. **e,** Number of raw reads vs. number of aligned reads in each human LR-RNA-seq library. **f,** Median length of each raw read vs. median length of the aligned portion of each read in each human LR-RNA-seq library. **g,** Read length profiles of post-TALON reads from LR-RNA-seq data split by tissue or cell line designation and polyA transcript length profile from GENCODE v40. **h,** Percentage of GENCODE vM25 polyA genes by gene biotype detected in at least one ENCODE mouse LR-RNA-seq library at >0 TPM, >=1 TPM, and >=100 TPM. **i,** Number of samples that each GENCODE vM25 gene is detected >= 1 TPM in the ENCODE mouse LR-RNA-seq dataset. **j,** Read length profiles of post-TALON reads from mouse LR-RNA-seq data split by tissue or cell line designation and polyA transcript length profile from GENCODE vM25.

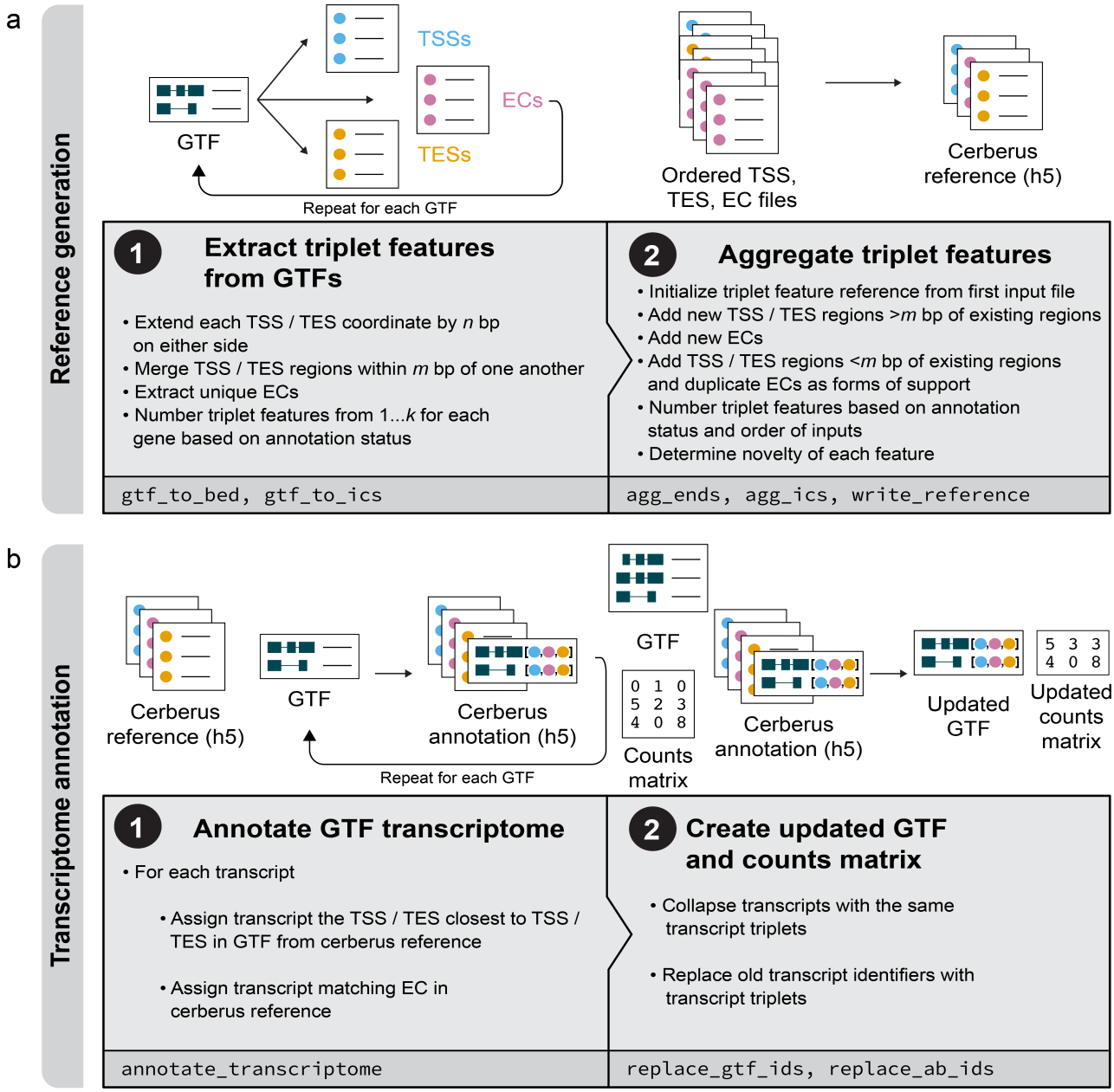

Figure S4. Overview of Cerberus processing of transcriptomes and triplet features. a, Workflow for generating a Cerberus reference: a collection of TSSs, ECs, and TESs (triplet features) sourced from various inputs. **b,** Workflow for generating a Cerberus transcriptome annotation, which assigns each transcript in a GTF a set of triplet features (TSS, EC, TES) from the Cerberus reference.

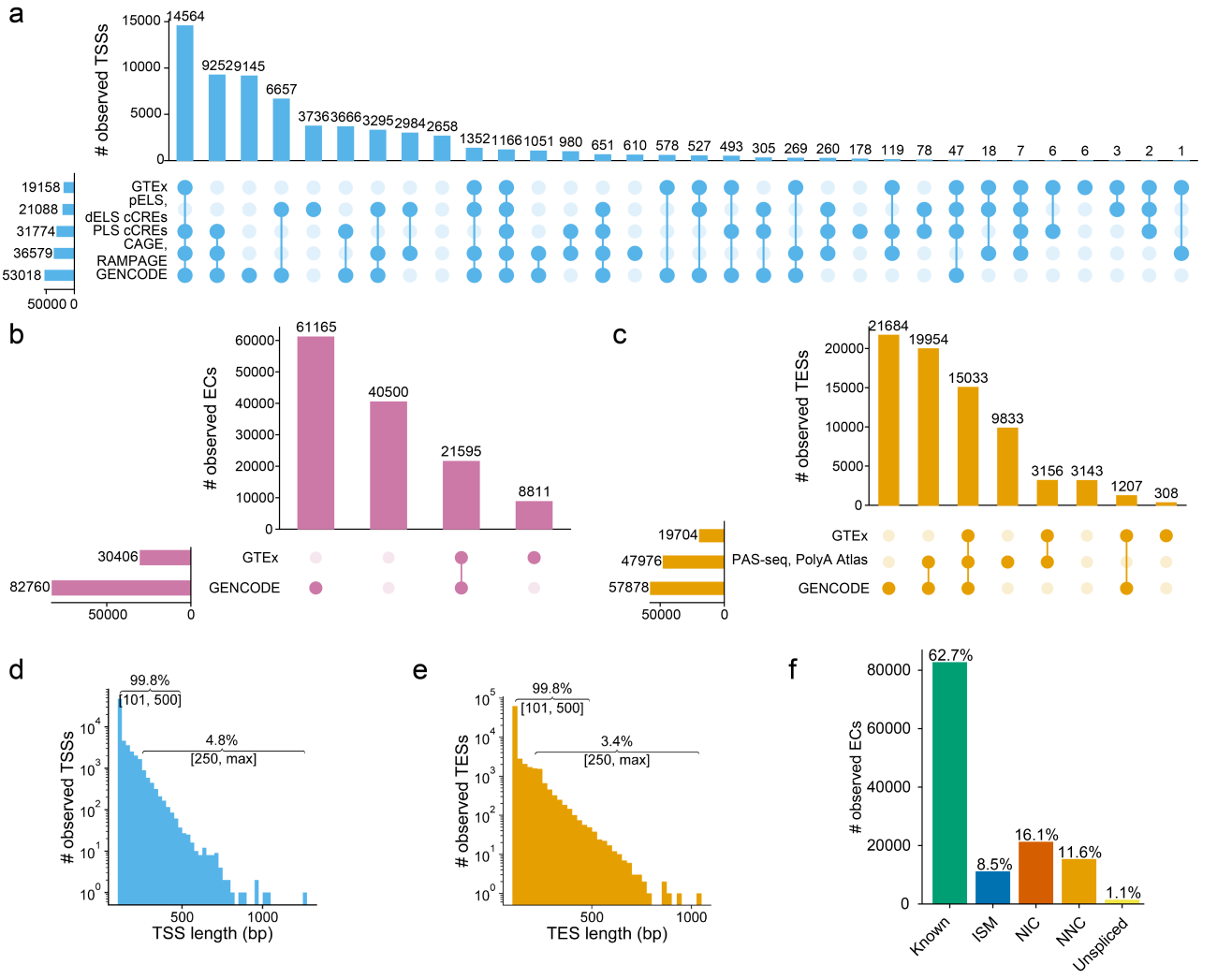

Figure S5. Characterization of observed triplet features from human LR-RNA-seq. a-c,Upset plots showing sources that overlap observed triplet features derived from human ENCODE LR-RNA-seq for **a,** TSSs **b,** ECs **c,** TESs. **d-e,** Lengths of observed **d,** TSSs **e,** TESs derived from human ENCODE LR-RNA-seq. **f** Novelty of unique ECs detected >= 1 TPM from polyA genes in human ENCODE LR-RNA-seq.

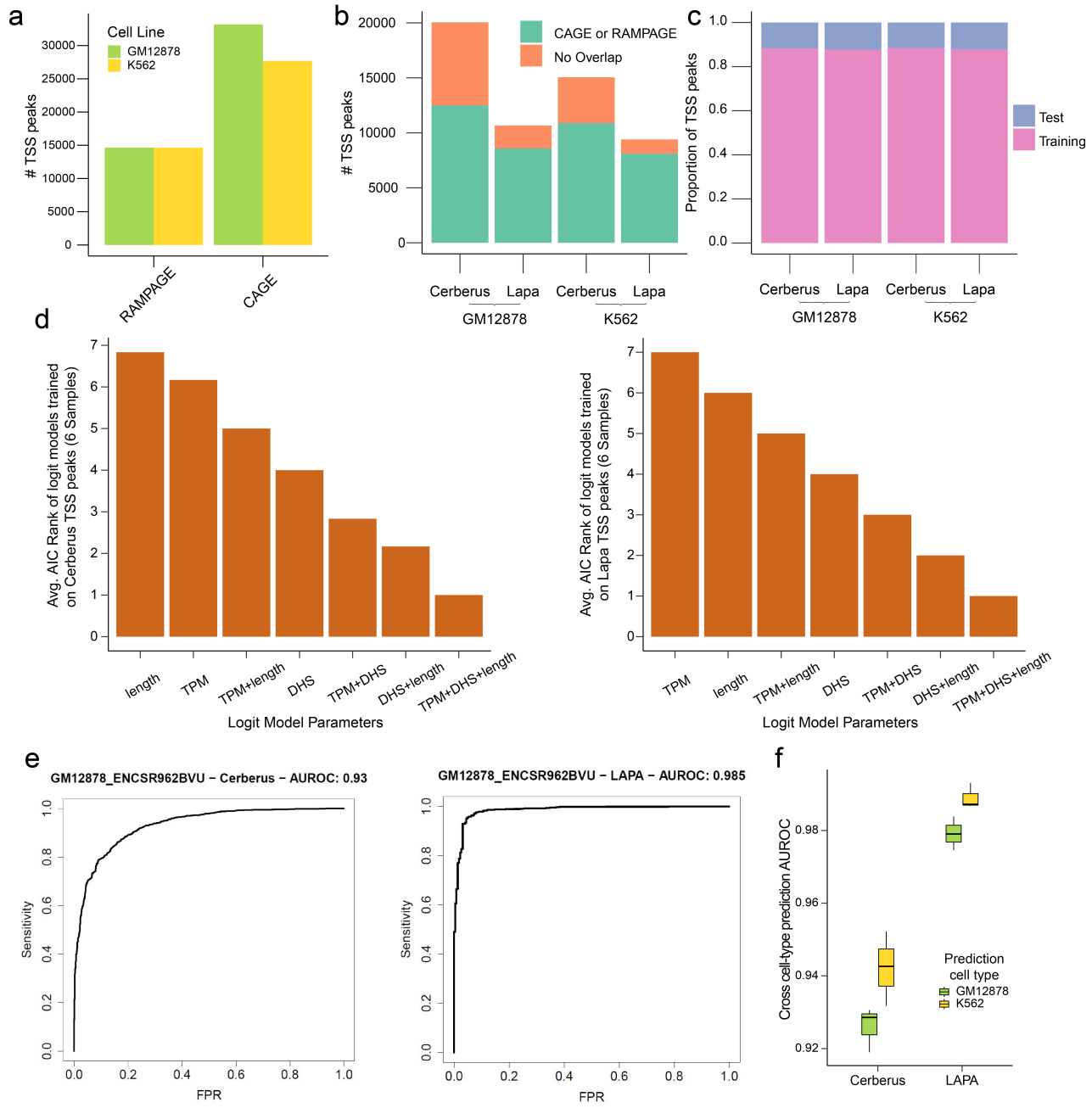

Figure S6. Machine learning models predict support for long-read TSS peaks by other TSS assays in a cross-cell type manner. a, Number of long-read RNA-seq TSS peaks called by RAMPAGE and CAGE. **b,** Number of long-read TSS peaks called by Cerberus or LAPA supported by RAMPAGE or CAGE assays in GM12878 and K562 long-read experiments gm12878_3 and k562_2. **c,** Fraction of peaks used for the test (chr 2 and chr3) and training sets (all other chromosomes) in K562 and GM12878 long-read experiments. **d,** Akaike Information Criterion (AIC) values for logistic regression models trained using different sets of parameters. For each experiment, the AIC values for the 7 training settings have been ranked. The y-axis is the average ranking of each model over all GM12878 and K562 long-read TSS peaks from Cerberus (left) and LAPA (right), where logit[overlap TPM + DHS + peak_length] is the best model (i.e with the lowest AIC). **e,** Same-cell type ROC curves for logit[overlap TPM + DHS + peak_length]. Models tested on chr2 chr3 and trained on other chromosomes in the same cell line. **f,** Cross-cell type logit[overlap TPM + DHS + peak_length]. Distribution of AUROC values for long-read TSS experiments. Ex: To predict if a long-read TSS peak in K562 overlaps with a region in K562 RAMPAGE and CAGE in a cross-cell type manner, a model is trained on a GM12878.

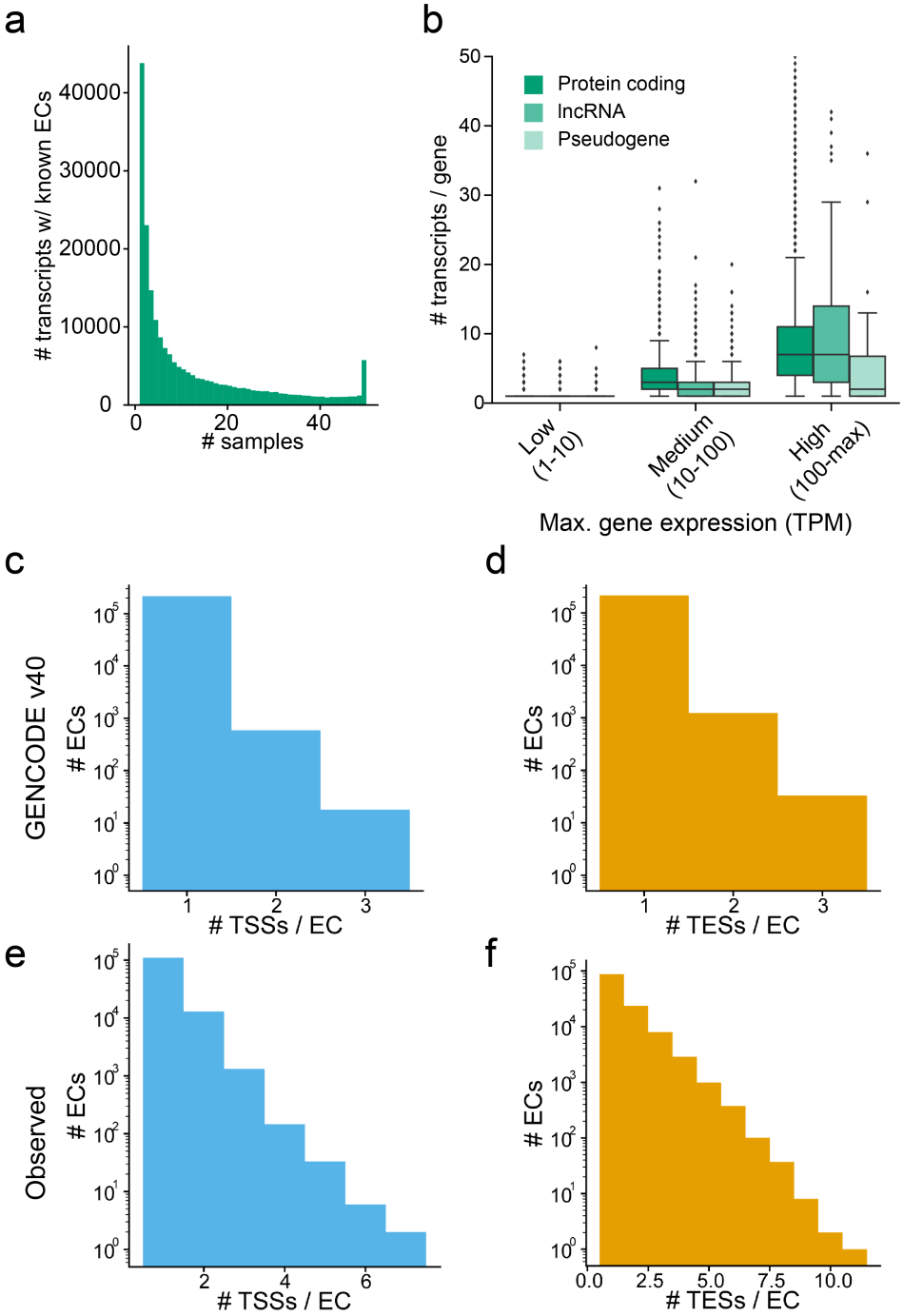

Figure S7. Characterization of observed transcripts from human LR-RNA-seq. a, Number of samples in which each transcript with a known EC is detected >= 1 TPM in the human ENCODE LR-RNA-seq dataset. **b,** Boxplot of, for the sample where a gene is most highly expressed, the number of transcripts expressed in that sample versus the TPM of the gene in that sample; split by gene biotype and gene expression bin. **c-f,** Number of unique TSSs or TESs per EC with at least 2 exons from transcripts **c-d,** annotated to polyA genes in GENCODE v40, **e-f,** detected >= 1 TPM from polyA genes in human ENCODE LR-RNA-seq.

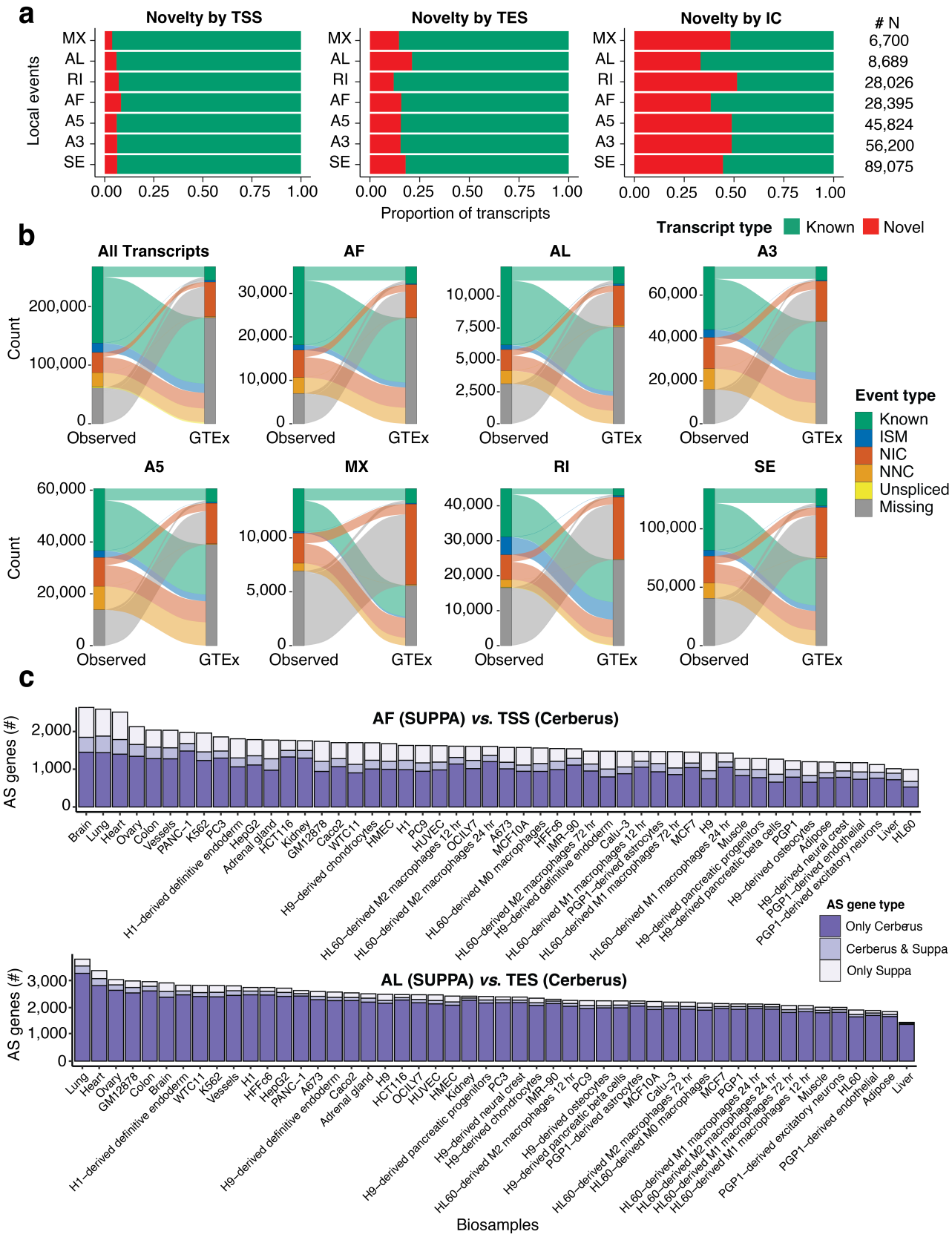

Figure S8. Alternative splicing (AS) detection by SUPPA and Cerberus. a, Barplots showing, for each type of local AS event detected by SUPPA (y axis), the proportion of observed known and novel transcripts identified by Cerberus (x axis), based on novel events at the TSS, TES, or EC. **b,** Sankey plots showing the number of transcripts classified as Known, Novel In Catalog (NIC), Novel Not in Catalog (NNC), Unspliced, or Missing by Cerberus and GTEx. In the first panel we show the numbers for all transcripts, while in the rest of the panels we focus on transcripts undergoing specific types of AS events. **c,** Barplot showing, for each LR-RNA-seq sample (x axis), the number of observed AS genes identified by both Cerberus and SUPPA (pink), only by Cerberus (dark gray), or only by SUPPA (light gray). In the upper panel we compare Alternative-First (AF)-AS genes by SUPPA and TSS-AS genes by Cerberus. In the lower panel we compare Alternative-Last (AL)-AS genes by SUPPA and TES-AS genes by Cerberus.

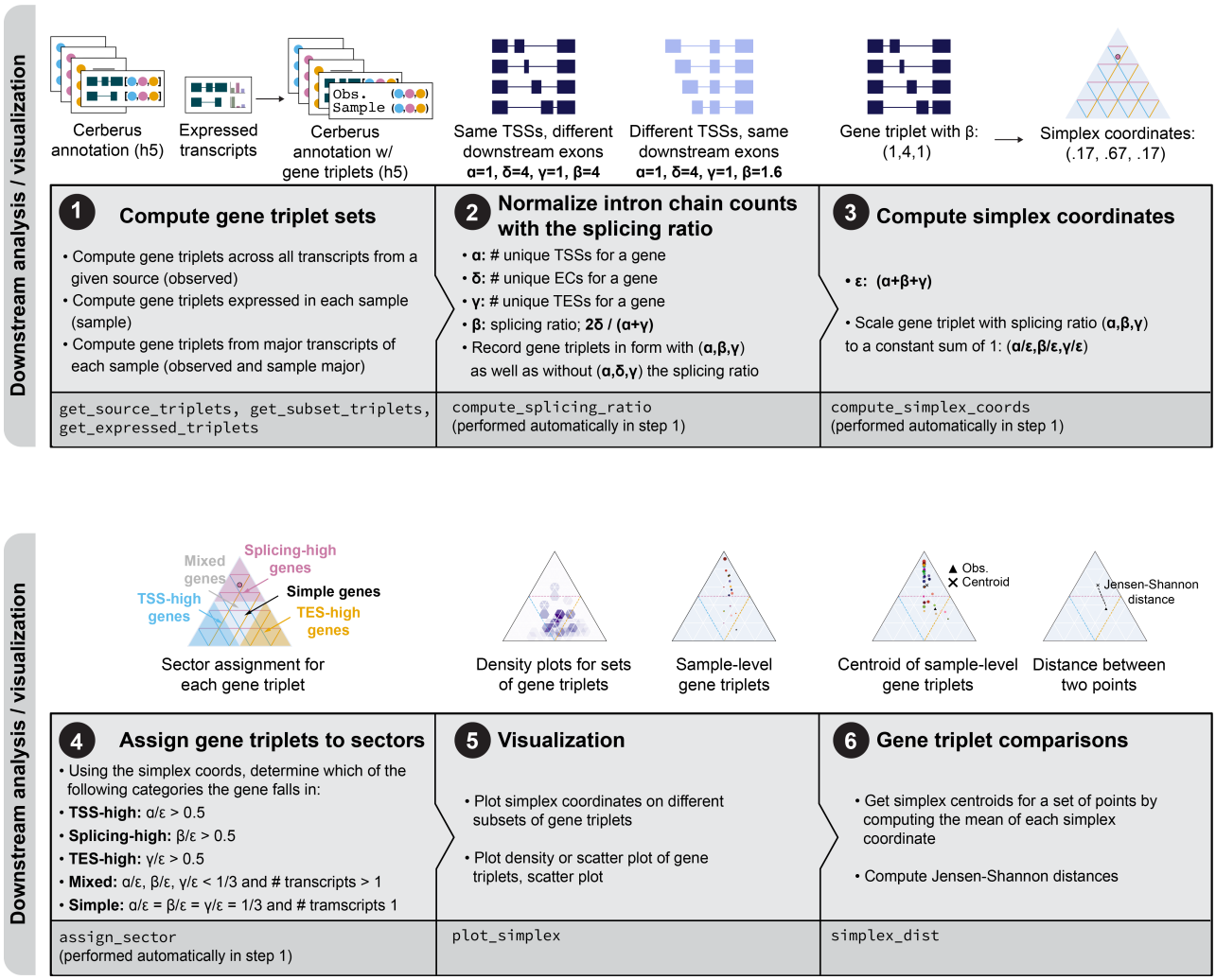

Figure S9. Overview of gene triplet based downstream analysis and visualization with Cerberus.

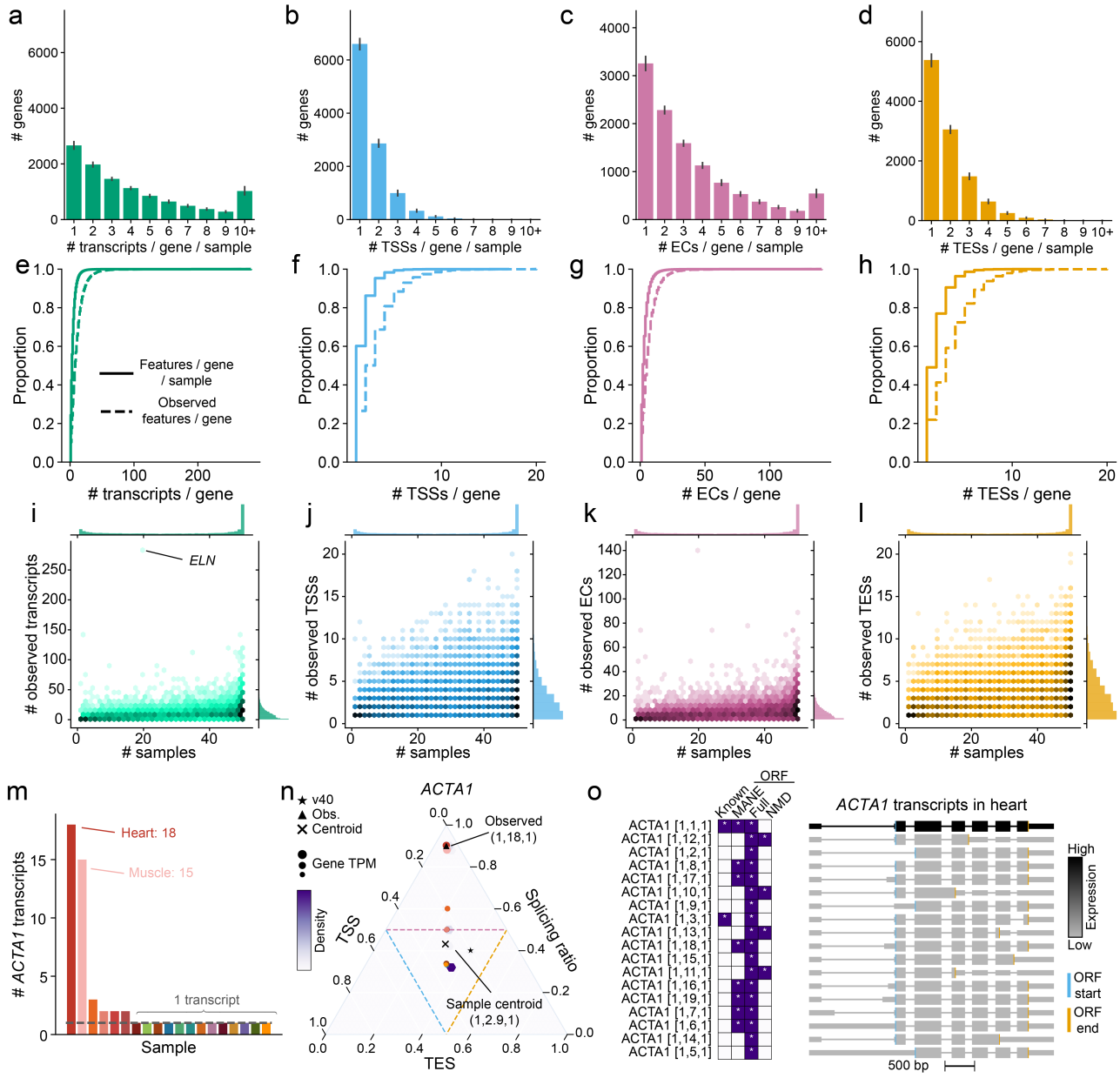

Figure S10. Uncovering sample-specific behavior of triplet features by comparing observed and sample-level gene triplets. a-d, Number of triplet features or transcripts detected per gene per sample for **a,** transcripts **b,** TSSs **c,** ECs **d,** TESs. **e-h,** Number of triplet features or transcripts per gene per sample and observed overall showing the proportion of the distribution that comes from each number of **e,** transcripts **f,** TSSs **g,** ECs **h,** TESs. **i-l,** Number of observed overall triplet features or transcripts per gene versus the number of samples each gene is expressed in for **i,** transcripts **j,** TSSs **k,** ECs **l,** TESs. **m,** Number of *ACTA1* transcripts expressed in each sample. **n,** Gene structure simplex for *ACTA1*. Gene triplets with splicing ratio for the overall observed and sample-level centroid labeled. Simplex coordinates for the GENCODE v40, observed set, and centroid of the samples also shown for *ACTA1*. **o,** Browser models of transcripts of *ACTA1* expressed >= 1 TPM in heart colored by expression level in TPM.

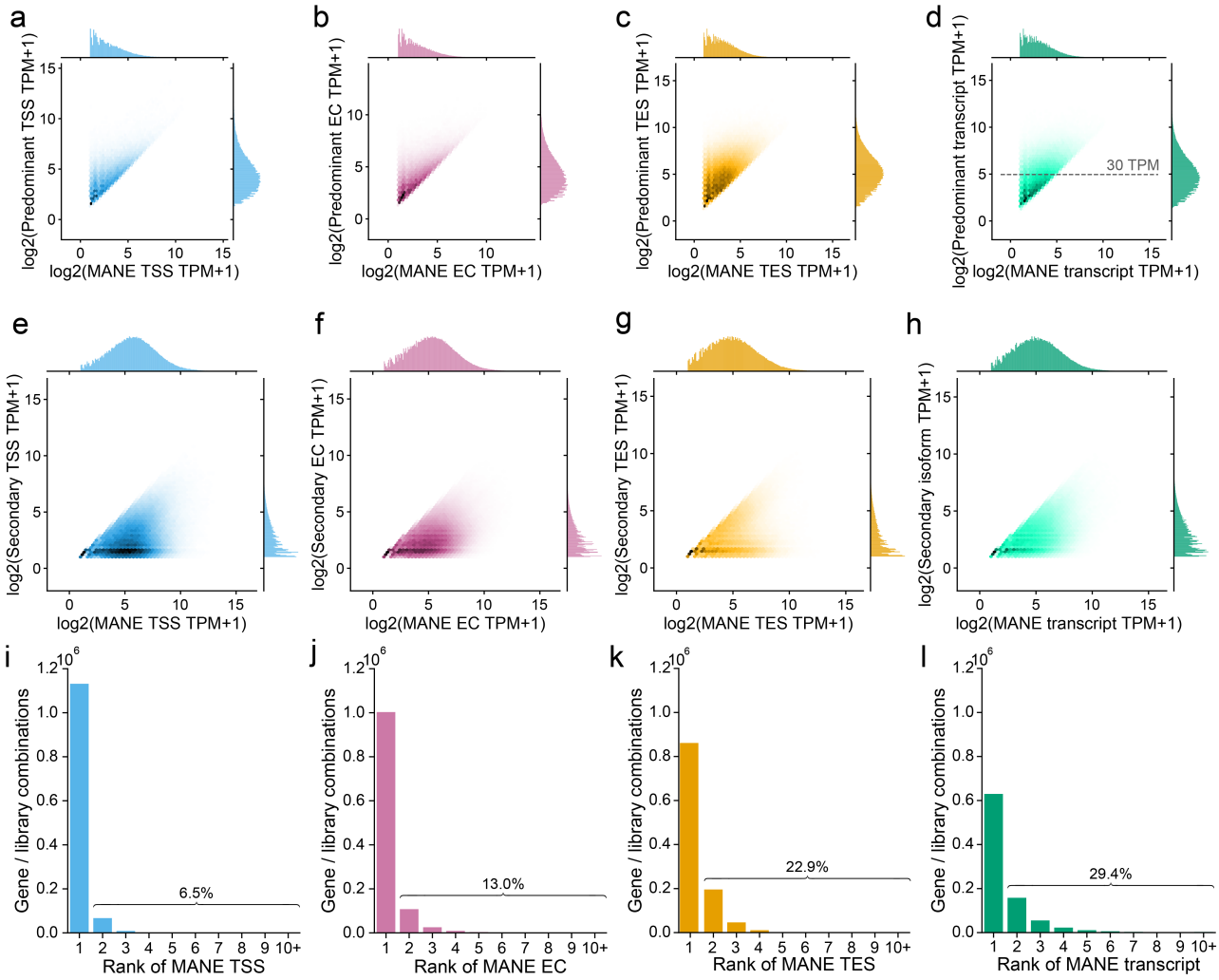

Figure S11. Rank and expression of predominant and MANE transcripts. a-d, For protein coding genes with MANE transcripts from GENCODE v40 where the predominant transcript or triplet feature is not the MANE transcript or triplet feature but is still expressed, expression of the predominant vs. the expression of the MANE **a,** TSS **b,** EC **c,** TES **d,** transcript. **e-h,** For genes where the MANE triplet feature or transcript is the predominant one and a secondary triplet feature or transcript is expressed, expression of the secondary vs. MANE **e,** TSS **f,** EC **g,** TES **h,** transcript. **i-l,** Rank of MANE **i** TSS **j,** EC **k,** TES **l,** transcript in each library where it is expressed.

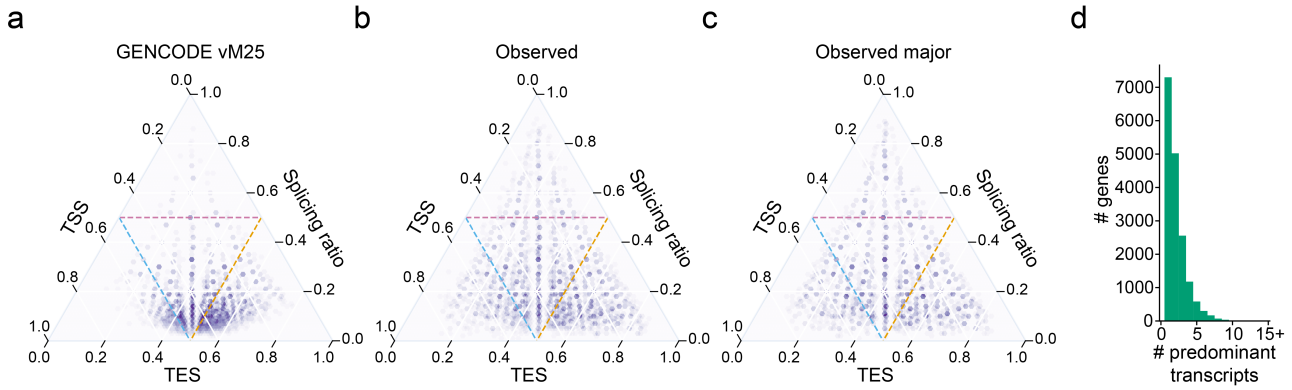

Figure S12. Rank and expression of predominant and MANE transcripts. a-c, Gene structure simplices for the transcripts from protein coding genes that are **a,** annotated in GENCODE vM25 where the parent gene is also detected in our mouse LR-RNA-seq dataset, **b,** the observed set of transcripts, those detected >= 1 TPM in the mouse ENCODE LR-RNA-seq dataset, **c,** the observed major set of transcripts, the union of major transcripts from each sample detected >= 1 TPM in the mouse ENCODE LR-RNA-seq dataset. **d,** Number of unique predominant transcripts detected >= 1 TPM across samples per gene.